Algorithmic statistics and useful information

Nikolay Vereshchagin[1]

[1]Moscow State University

VAI 2015

Algorithmic statistics

Question:

Assume that a bit string x (data) and a statistical hypothesis P (a probability distribution over strings) are given; when do we consider P a good "explanation" for x?

Algorithmic statistics

Question:

> *Assume that a bit string x (data) and a statistical hypothesis P (a probability distribution over strings) are given; when do we consider P a good "explanation" for x?*

A technical restriction: we will consider only uniform distributions over finite sets as statistical hypotheses.

Repeating the Question:

> *Assume that a bit string x (data) and a set A containing x (a model) are given; when do we consider A a good "explanation" for x?*

A short answer

Example

1. A random string
   $x = 0111110001111101001000000100011000011001000001011101$
   of length 100.

   Good explanation: $A = \{0,1\}^{100}$ (the set of all 100-bit sequences)
   Bad explanation: $B = \{x\}$.

2. $y = 0000000000000000000000000000000000000000000000000000$

   Good explanation: $C = \{y\}$.
   Bad explanation: $A = \{0,1\}^{100}$.

A short answer

Example

1. A random string
   $x = 011111000111110100100000010001100001100100001011101$
   of length 100.

   Good explanation: $A = \{0,1\}^{100}$ (the set of all 100-bit sequences)
   Bad explanation: $B = \{x\}$.

2. $y = 000000000000000000000000000000000000000000000000000$

   Good explanation: $C = \{y\}$.
   Bad explanation: $A = \{0,1\}^{100}$.

A model $A \ni x$ is a good "explanation" for x if

1. A is simple,
2. the string x is a random (typical) element of A.

Kolmogorov complexity

$C(x)$, $C(x|y)$, $C(A)$, $C(x, y)$, $C(A|x)$, $C(x|A)$, $C(x|y, z)$ etc.

$C(0111110001111101001000000100011000011) \approx n$.
$C(0000000000000000000000000000000000000) \approx \log_2 n$.

Definition

x is *independent on y* if $C(x|y) \approx C(x)$.

Notice that $C(x|y) \leq C(x)$ for all x, y.

$C(x) - C(x|y)$ is called the *information in y about x*.

Theorem (Symmetry of information, Kolmogorov–Levin)

$$C(x) + C(y|x) = C(y) + C(x|y) = C(x, y).$$

Simple sets

Definition

A set $A \ni x$ is *simple explanation of* x if $C(A)$ is small compared to the length of x.

Convention: We adopt logarithmic accuracy, that is, *small* means of order $O(\log |x|)$ where x is a data string.

Simple sets

Definition

A set $A \ni x$ is *simple explanation of* x if $C(A)$ is small compared to the length of x.

Convention: We adopt logarithmic accuracy, that is, *small* means of order $O(\log |x|)$ where x is a data string.

Example

Both sets

$$\{0,1\}^n \text{ and } \{\underbrace{0000000000000000000000000000}_{n \text{ times}}\}$$

are simple. They have the same complexity, which is at most $\log_2 n + O(1)$.

Randomness deficiency

Definition

Notice that $C(x|A) \leq \log_2 |A|$ for all $A \ni x$. A string x is *a random element of a set $A \ni x$* if

$$C(x|A) \approx \log_2 |A|.$$

The quantity

$$\log_2 |A| - C(x|A)$$

is called *the randomness deficiency of x in A*.

Randomness deficiency

Definition

Notice that $C(x|A) \leq \log_2 |A|$ for all $A \ni x$. A string x is *a random element of a set $A \ni x$* if

$$C(x|A) \approx \log_2 |A|.$$

The quantity

$$\log_2 |A| - C(x|A)$$

is called *the randomness deficiency of x in A*.

Example

1. Assume that $C(x) \approx |x| = n$. Then x is a random element of $\{0,1\}^n$.
2. The string 0000000000000000000000 consisting of n zeros is not a random element of the set $\{0,1\}^n$.
3. However the string 0000000000000000000000 is a random element of the set $\{0000000000000000000000\}$.

Stochastic strings

Definition (Kolmogorov' 1983)

A string x is called *stochastic* there is a simple set $A \ni x$ such that x is a random element of A.
Otherwise x is called *non-stochastic.*

Example

Assume that $C(x) \approx |x| = n$. Then x is stochastic, witnessed by the set $A = \{0,1\}^n$:

$$C(A) \approx 0, \qquad C(x|A) \approx n = \log_2 |A|.$$

Theorem (Shen' 1983)

There are non-stochastic strings.
More specifically, for all n there is a string of length n whose randomness deficiency is at least $n/3$ in every set of complexity less than $n/3$.

Useful information: identifying noise

The Model Example. Let y be any string and z a random string independent on y,

$$C(z|y) \approx C(z) \approx |z|.$$

Let $x = (y, z)$. Then z is the noise in x. Thus all useful information from x is inside y.

Definition

A pair y, z *identifies noise in* x if

1. x is equivalent to the pair y, z
2. and z is a random string independent on y.

In this case we say that z is *a noise in* x.

Definition

A string x *the same or more information* than a string y, $x \to y$, if $C(y|x) \approx 0$. Strings x and y are (informational) equivalent, $x \leftrightarrow y$, if $C(x|y) \approx 0$ and $C(y|x) \approx 0$.

Useful information: two part codes

Definition (a reminder)

A pair y, z *identifies noise in* x if

1. $(y, z) \leftrightarrow x$,
2. and $C(z|y) \approx C(z) \approx |z|$.

Lemma

A pair y, z identifies noise in x iff

$$(y, z) \to x \qquad \textit{and} \qquad C(y) + |z| \approx C(x).$$

The pair (y, z) is called a **two part code** *for x, where y is the* **model** *and z is the* **data-to-model** *code for x.*

Proof.

$$C(x) \leq C(y, z) = C(y) + C(z|y) \leq C(y) + |z|.$$

□

The “naive” approach fails

Lemma

The empty string captures useful information from any string x.

Proof.

The pair (the empty string, the shortest program for x) identifies noise in x. $\square$

Question: What’s wrong with this “naive” approach?
Answer: The time to transform (y, z) to x may be huge. It may be not bounded by any total computable function.

Useful information: Koppel's approach (1988)

Koppel considered two part codes of the form:
(a total computable function f, a string z with $f(z) = x$).

Definition (Koppel)

The *sophistication* of a string x is the minimal length of a **total** program p such that for some string z

1. $p(z) = x$,
2. $|p| + |z| \approx C(x)$

(in which case the pair p, z identifies noise in x).

Useful information: Kolmogorov's approach (1974, unpublished)

Kolmogorov considered two part codes of the form:
(a finite set A, the index of x in A).

Definition (Kolmogorov)

A finite set A is called a *sufficient statistic for* x if

1. $x \in A$,
2. $C(A) + \log_2 |A| \approx C(x)$

(in which case the pair (A, the index of x in A) identifies noise in x).

Lemma

A set A is a sufficient statistic for x iff x is a random element in A and $C(A|x) \approx 0$ (i.e. A has no redundant information).

Definition

A set A is called a *minimal sufficient statistic for* x if A is a sufficient statistic for x of minimal complexity.

Koppel's approach = Kolmogorov's approach

Kolmogorov $\Rightarrow$ Koppel:
$A \ni x \Rightarrow (p, z)$
where $z =$ (the index of x in A)
and p is a shortest program mapping i to ith element of A (and to the empty string, say, if $i > |A|$).

Koppel's approach = Kolmogorov's approach

Kolmogorov $\Rightarrow$ Koppel:
$A \ni x \Rightarrow (p, z)$
where $z =$ (the index of x in A)
and p is a shortest program mapping i to ith element of A (and to the empty string, say, if $i > |A|$).

Koppel $\Rightarrow$ Kolmogorov:
p, z with $p(z) = x \Rightarrow$ the set $A = \{p(z') \mid |z'| = |z|\}$.

Koppel's approach = Kolmogorov's approach

Kolmogorov $\Rightarrow$ Koppel:
$A \ni x \Rightarrow (p, z)$
where $z =$ (the index of x in A)
and p is a shortest program mapping i to ith element of A (and to the empty string, say, if $i > |A|$).

Koppel $\Rightarrow$ Kolmogorov:
p, z with $p(z) = x \Rightarrow$ the set $A = \{p(z') \mid |z'| = |z|\}$.

Example (trivial sufficient statistics)

Let x be any string.
The set $A = \{x\}$ is a sufficient statistic for x.
Here $C(A) \approx C(x)$.

Example (sufficient statistics for stochastic strings)

Assume that x is stochastic, witnessed by the set A.
Then A is a sufficient statistic for x. Indeed,

$$C(A) + \log_2 |A| \approx \log_2 |A| \approx C(x|A) \approx C(x).$$

Example (The Model Example)

Let y be a non-stochastic string and z a random string independent on y.
Let $x = (y, z)$.
Then the set

$$A = \{(y, z') \mid |z'| = |z|\}$$

is a sufficient statistic for x.

Sophistication, minimal sufficient statistics and useful information

Definition

The *amount of useful information in a data string* x is its sophistication ($=$ the complexity of a minimal sufficient statistic for x).

Example

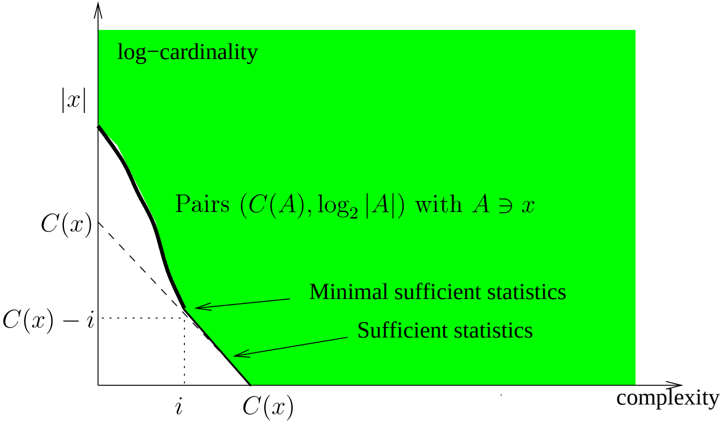
Stochastic strings (and only they) have no useful information.

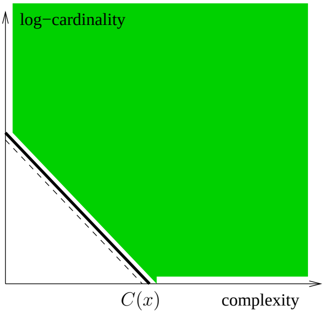
Minimal sufficient statistics

Lemma

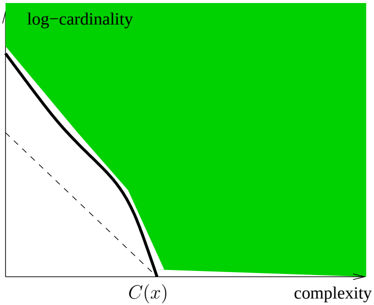
If x has a sufficient statistic of complexity $i < C(x)$ then it has a sufficient statistic of every complexity in the interval $[i, C(x)]$.

The profile of x:





Stochastic strings.



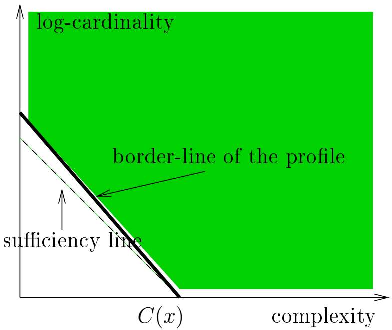
Highly non-stochastic strings

Definition

A string x is *highly non-stochastic* if the complexity of any its sufficient statistic is close to $C(x)$. That is, all information in x is useful.

Caution

We ignore terms of order $\log |x|$ (where x is the data string). As a result the definition of a minimal sufficient statistics becomes quite vague.
Example:



Convention: We will consider only strings x for which the border-line of the profile of x either does not leave the sufficiency line or leaves it at an angle that is larger than 45 degrees.

Good news

Theorem (V, Vitányi' 2002)

Highly non-stochastic strings exist. Moreover, for any given profile satisfying obvious constraints there is a string having that profile.

Theorem (V' 2009)

Let $x = (y, z)$ where y is highly non-stochastic and z is a random string independent on y. Then the set

$$A = \{(y, z') \mid |z'| = |z|\}$$

is a minimal sufficient statistic for x. In other words, the amount of useful information in x is $C(y)$ (and not less).

Surprisingly good news

Assume that we a given a data string x and a "threshold" β. Consider the following two tasks.

1. Task 1: Minimize $C(A)$ under the constraints

$$x \in A, \qquad \log_2 |A| - C(x|A) \leq \beta.$$

   This is the task of finding a good statistical explanation of the given data.
   If the complexity of optimal solution is less than α then the string x is called α, β-*stochastic* (Kolmogorov).

2. Task 2: Minimize $C(A)$ under the constraints

$$x \in A, \qquad \log_2 |A| + C(A) \leq C(x) + \beta.$$

   This is the task of denoising the data.

All admissible solutions for the second task are admissible solutions for the first task but not the other way around. (Example: $\{0,1\}^n \setminus \{y\}$ as a model for a random string x independent on y.)

Theorem (V, Vitányi' 2002)

These tasks are equivalent: any optimal solution to the first task is an optimal solution to the second task and the other way around.

Bad news

Theorem (Gács, Tromp, Vitányi' 1998, V, Vitányi' 2002)

1. *If A and B are minimal sufficient statistics for x then $A \leftrightarrow B$.*

Bad news

Theorem (Gács, Tromp, Vitányi' 1998, V, Vitányi' 2002)

1. *If A and B are minimal sufficient statistics for x then $A \leftrightarrow B$.*
2. *Moreover, if A is a minimal sufficient statistic for x and its complexity is i then*

$$A \leftrightarrow \Omega_i$$

where Ω_i denotes the number of strings of complexity at most i.

Bad news

Theorem (Gács, Tromp, Vitányi' 1998, V, Vitányi' 2002)

1. *If A and B are minimal sufficient statistics for x then $A \leftrightarrow B$.*
2. *Moreover, if A is a minimal sufficient statistic for x and its complexity is i then*
$$A \leftrightarrow \Omega_i$$
where Ω_i denotes the number of strings of complexity at most i.
3. *Moreover, there is a "universal" family of models $\{S_{ik} \mid i, k \in \mathbb{N}, i \leq k\}$ such that*
$$S_{ik} \leftrightarrow \Omega_i, \qquad C(S_{ik}) = i, \quad \log_2 |S_{ik}| = k - i$$
and for every string x there is a minimal sufficient statistic S_{ik} for x with $k \approx C(x)$.

What is wrong with the approach of Koppel and Kolmogorov?

What is wrong with the approach of Koppel and Kolmogorov?

1. It seems that our definition of "having the same information" is too broad, we assumed that u and v are informational equivalent if both $C(u|v)$ and $C(v|u)$ are negligible.
   Under this assumption every string x has the same information as its shortest program x^*.
   In the context of separating the information into a useful one and an accidental one, such an assumption is certainly misleading. Indeed, for any string x we have $x \leftrightarrow x^*$. The string x^* is always stochastic while x may be highly non-stochastic.

What is wrong with the approach of Koppel and Kolmogorov?

1. It seems that our definition of "having the same information" is too broad, we assumed that u and v are informational equivalent if both $C(u|v)$ and $C(v|u)$ are negligible.
   Under this assumption every string x has the same information as its shortest program x^*.
   In the context of separating the information into a useful one and an accidental one, such an assumption is certainly misleading. Indeed, for any string x we have $x \leftrightarrow x^*$. The string x^* is always stochastic while x may be highly non-stochastic.
2. However, even if we adopt a more restrictive definition of informational equivalence, the universal models S_{ik} discredit the approach.

Questions:

1. Is there a natural more restrictive definition of informational equivalence?
2. Is it possible to restrict the class of sufficient statistics so that to ban universal models S_{ik} while keeping “natural” models like those from the examples?

Questions:

1. Is there a natural more restrictive definition of informational equivalence?
2. Is it possible to restrict the class of sufficient statistics so that to ban universal models S_{ik} while keeping "natural" models like those from the examples?

Answers:

1. Again we neglect computation time. We should think that x and y are informational equivalent if there are short programs mapping x to y and back in a "reasonable" time.
2. We will try.

Total conditional complexity

Definition (Total conditional complexity)

$CT(x|y)$
$= \min\{|p| : p(y) = x,\ p(y') \text{ halts for all } y'\}$.

Total conditional complexity

Definition (Total conditional complexity)

$CT(x|y)$
$= \min\{|p| : p(y) = x,\ p(y') \text{ halts for all } y'\}$.

Lemma

For all n there is a string x of length n with

$$CT(x|x^*) = \Omega(n)$$

for all short programs x^* for x.

Total conditional complexity

Definition (Total conditional complexity)

$CT(x|y)$
$= \min\{|p| : p(y) = x,\ p(y') \text{ halts for all } y'\}.$

Lemma

For all n there is a string x of length n with

$$CT(x|x^*) = \Omega(n)$$

for all short programs x^* for x.

Theorem (Bauwens, Makhlin, V, Zimand' 2013)

For all x there is a short program x^* for x with

$$CT(x^*|x) = O(\log n).$$

A finer approach to the definition of "having the same information"

Definition

Strings x and y are *strongly* (informational) equivalent, $x \Leftrightarrow y$, if $CT(x|y) \approx 0$ and $CT(y|x) \approx 0$.

Lemma

If $x \Leftrightarrow y$ then the profiles of x and y are close to each other.

Question: Is every minimal sufficient statistic **strongly** equivalent to some Ω_i?
Answer: Not any more!

Theorem (V' 2015)

There is a string and its minimal sufficient statistic A that is not strongly equivalent to Ω_i (for any i).

Restricting sufficient statistics: strong statistics

Definition

A is a *strong statistic for* x if $CT(A|x) \approx 0$.
A is a *good statistic for* x if A is a strong sufficient statistic for x.
Useful information in the strong sense $=$ minimal good statistic for x.

Remark. If A is sufficient statistic for x then $C(A|x) \approx 0$. However it may happen that $CT(A|x) \gg 0$.

Restricting sufficient statistics: strong statistics

Definition

A is a *strong statistic for* x if $CT(A|x) \approx 0$.
A is a *good statistic for* x if A is a strong sufficient statistic for x.
Useful information in the strong sense = minimal good statistic for x.

Remark. If A is sufficient statistic for x then $C(A|x) \approx 0$. However it may happen that $CT(A|x) \gg 0$.

Example

Let $x = (y, z)$ where y is highly non-stochastic string and z is a random strings independent on y. Then the set

$$\{(y, z') \mid |z'| = |z|\}$$

is a minimal good statistic for x.

Good statistics

Lemma

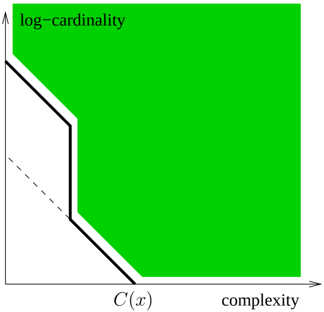
A is a good statistic for x iff x and the pair (A, the index of x in A) are strongly equivalent and the index of x in A is a random string independent on A.

Question: Are there indeed sufficient statistic that are not strong?
Answer: Yes!

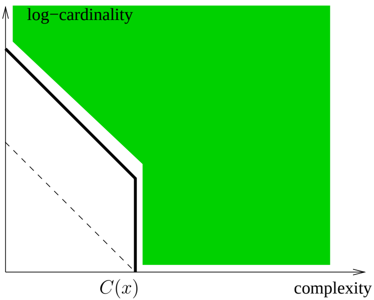
Strange strings

Theorem (on existence of strange strings, V'2012)

There is a string x whose profile and strong profile are far apart:



The profile of x



The strong profile of x

Banning universal models S_{ik}

Question: Does restricting to good statistics ban any S_{ik}?
Answer: Yes!

Theorem (Milovanov' 2015)

There is a string x that has strong minimal sufficient statistic but no model of the form S_{ik} is such a statistic.

Transforming one model into another one

Question: Assume that we have two sufficient statistics A and B for the same string x and $C(A) \geq C(B)$. Is it true that $C(B|A) \approx 0$?
Answer: No.

Example

Let

$$x = 011111000111110100100000100011000011$$

be a random string of length n. Let

$$A = \{\underbrace{*************}_{n/3}10100100000100011000011\}$$

and

$$B = \{011111000111110100\underbrace{*****************}_{n/2}\}$$

Uniqueness of minimal good statistic

Question: Assume now that A, B are sufficient statistics A and B for the same string x and B is *minimal*. Is it true that $C(B|A)$ is negligible?
In our example, any minimal sufficient statistic has a very small complexity, as x is stochastic, thus the answer is positive by trivial reasons.

Uniqueness of minimal good statistic

Question: Assume now that A, B are sufficient statistics A and B for the same string x and B is *minimal*. Is it true that $C(B|A)$ is negligible?
In our example, any minimal sufficient statistic has a very small complexity, as x is stochastic, thus the answer is positive by trivial reasons.

Answer: Yes. Moreover, if B is a strong statistic for x then the total complexity $CT(B|A)$ is negligible.

Theorem (V'2009)

Assume that B is a minimal sufficient statistic for x and A is a sufficient statistic for x. Then $C(B|A) \approx 0$. If, additionally, B is a strong statistic for x then $CT(B|A) \approx 0$.

Sufficient statistics for sufficient statistics

Lemma

Any minimal sufficient statistic for a non-stochastic object is highly non-stochastic.

Proof.

Let A be a minimal sufficient statistic for x and B minimal sufficient statistic for A. Assume that $C(B) \ll C(A)$.

Then we are able to construct a sufficient statistic A' for x with $C(A') \ll C(A)$ applying to B *the Lifting Procedure*:
Let

$$A' = \bigcup \{X \in B \mid |X| \approx |A|\}$$

Then $C(A') \approx C(B)$ and A' is a sufficient statistic for x. Indeed, we have

$$\log_2 |A'| \leq \log_2 |B| + \log_2 |A|.$$

Thus

$$C(A') + \log_2 |A'| \leq C(B) + \log_2 |B| + \log_2 |A| \approx C(x).$$

Step-wise denoising

Scenario:
there is a data string x such that there is a strong minimal sufficient statistic for x. Our goal is to denoise it, that is, to find such a statistic.

Assume that somebody performed a partial denoising of x obtaining a good model A for x and then another guy fully denoised A and gave us a minimal sufficient statistic D for A:

1. $x \xRightarrow{\text{Partial denoising}}$ a good statistic A for x
2. $A \xRightarrow{\text{Full denoising}}$ a minimal sufficient statistic D for A

Can we recover a minimal sufficient statistic for x from D? A natural idea is to apply the lifting to D. The complexity of resulting sufficient statistic B for x is $C(D)$ and B is a good statistic for x provided D is good.

Question: Is D a minimal sufficient statistic for x?

Step-wise denoising

Scenario:
there is a data string x such that there is a strong minimal sufficient statistic for x. Our goal is to denoise it, that is, to find such a statistic.

Assume that somebody performed a partial denoising of x obtaining a good model A for x and then another guy fully denoised A and gave us a minimal sufficient statistic D for A:

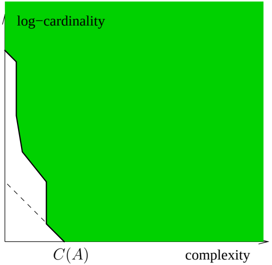
1. $x \overset{\text{Partial denoising}}{\Longrightarrow}$ a good statistic A for x
2. $A \overset{\text{Full denoising}}{\Longrightarrow}$ a minimal sufficient statistic D for A

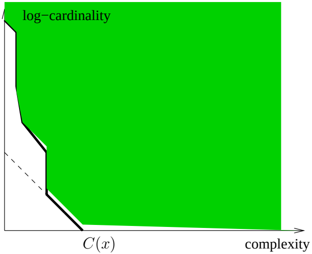
Can we recover a minimal sufficient statistic for x from D? A natural idea is to apply the lifting to D. The complexity of resulting sufficient statistic B for x is $C(D)$ and B is a good statistic for x provided D is good.

Question: Is D a minimal sufficient statistic for x?
Answer: Yes.

Theorem (V'2009)

If A is a good statistic for x then the complexities of minimal sufficient statistics for x and A are close. Moreover, the profiles of x and A look, as shown on the following figure:





Normal strings

Definition

A string x is called *normal* if its strong profile is close to its profile.

Question: Are normal non-stochastic strings rare?

Normal strings

Definition

A string x is called *normal* if its strong profile is close to its profile.

Question: Are normal non-stochastic strings rare?
Answer: No!

Theorem (on existence of normal strings, Milovanov' 2015)

For any given string x there is a normal string having the same profile (with $O(\sqrt{|x|})$ accuracy).

Normal strings

Definition

A string x is called *normal* if its strong profile is close to its profile.

Question: Are normal non-stochastic strings rare?
Answer: No!

Theorem (on existence of normal strings, Milovanov' 2015)

For any given string x there is a normal string having the same profile (with $O(\sqrt{|x|})$ accuracy).

Question: Assume that we denoised a normal string x and obtained a minimal good statistic A for x. Is A always normal?

Normal strings

Definition

A string x is called *normal* if its strong profile is close to its profile.

Question: Are normal non-stochastic strings rare?
Answer: No!

Theorem (on existence of normal strings, Milovanov' 2015)

For any given string x there is a normal string having the same profile (with $O(\sqrt{|x|})$ accuracy).

Question: Assume that we denoised a normal string x and obtained a minimal good statistic A for x. Is A always normal?
Answer: Yes!

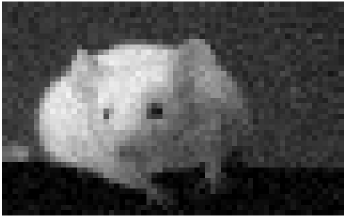
Theorem (Normality is hereditary, Milovanov' 2015)

Assume that x is a normal string and A is a minimal good statistic for x. Then A is normal as well (with $O(\sqrt{|x|})$ accuracy).

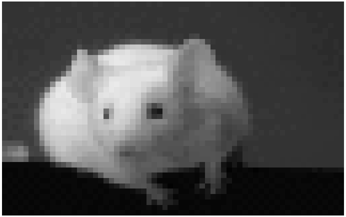
Some applications.

Denoising a real data (de Rooij, Vitányi' 2012)

Noisy mouse (64×40 pixels):



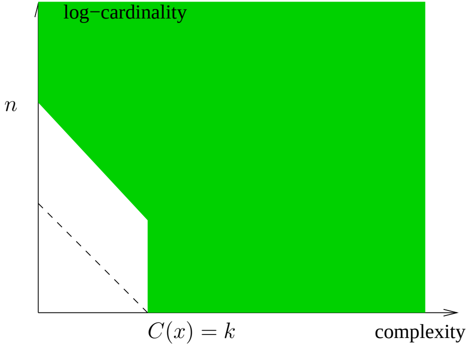
Denoised mouse:



List decoding from erasures (Milovanov' 2015)

Definition

A string of length n and complexity k is called *anti-stochastic* if it has the smallest possible profile for strings of that length and complexity:



Theorem (Holographic property of anti-stochastic strings)

Every anti-stochastic string x of length n and complexity k can be restored from any string $\tilde{x}$ obtained from x by erasing any $n-k$ its bits (erased symbols are replaced by $*$) by a program of length $O(\log n)$. That is,
$C(x|\tilde{x}) = O(\log n)$.
Such strings are called n, k-holographic.

Corollary

There are about 2^k n, k-holographic strings. They thus form a code of rate k/n that is capable to correct $n-k$ erasures by list decoding with list size $poly(n)$.

Thank you.